# Labor Economics - Empirical Assignment

*By* Evgeniya Dubinina, Martin Kosík, Lukáš Supik and Jan Žemlička

1) We merged the data in R programming language using the `left_join()` function from the `dplyr` package. The correctness of the results was checked manually.

2) Only 3 variables contain missing data: main activity of the company (c06), the number of working hours per day (c13), and average monthly earnings (c14). The number of missing data points for each variable is 18, 22, and 101, respectively. Furthermore, 2 observations had negative average monthly earnings. We decided to drop the observations with missing data or negative earnings. This results in reduction of the number of observation from 3837 to 3700. Since the dropped data makes up relatively small share of the full dataset (less than 4%), it is unlikely to substantially affect the results of the subsequent analyses. In addition, we also excluded observations with negative experience (11 in total) which are likely result of some error.

3) The table 1 shows the descriptive statistics for the variables of interest. The "exper" variable refers to experience which was calculated as exper = age - schooling_years -6, "exper_sq" is its square, "ln_y" is a natural logarithm of earnings, and "prague" is a dummy variable that equals one if the respondent lives in Prague and zero otherwise. We see that average person in the sample has around 20 years of education, earns 7 484 CZK per month, and has around 13 years of schooling. Nevertheless there is substantial variability around these averages.

Table 1—Descriptive statistics

| | variable | n | Mean | Std. Dev. | Median | Min. | Max. |
|---|---|---|---|---|---|---|---|
| 1 | any_children | 3689 | 0.40 | 0.49 | 0.00 | 0.00 | 1.00 |
| 2 | earnings | 3689 | 7483.60 | 4188.35 | 6666.67 | 300.00 | 50000.01 |
| 3 | exper | 3689 | 20.61 | 11.62 | 21.00 | 0.00 | 74.00 |
| 4 | exper_sq | 3689 | 559.52 | 523.20 | 441.00 | 0.00 | 5476.00 |
| 5 | ln_y | 3689 | 8.80 | 0.49 | 8.80 | 5.70 | 10.82 |
| 6 | prague | 3689 | 0.12 | 0.32 | 0.00 | 0.00 | 1.00 |
| 7 | schooling_years | 3689 | 12.58 | 2.38 | 12.00 | 8.00 | 25.00 |
| 8 | urate | 3689 | 0.03 | 0.02 | 0.03 | 0.00 | 0.08 |

4) a) The results of the basic Mincerian specification are provided in table 2, column 1.

b) We chose the logarithm of average monthly earnings (variable c14 in the dataset) as our left-hand side variable. The reason why earnings might be preferable to hourly wage is that it we can then include entrepreneurs and self-employed into the sample (since they do not necessarily receive fixed hourly wage, but they do have some monthly earnings).

We restricted the sample to include only men working full-time (6 to 10 hours per day), so that we would not have to deal with issues arising when comparing part-time and full-time workers.

We do observe some significant differences between men who work full-time those who do not. First, men who work full-time have on average 0.58 fewer years of schooling than men who do not (SE of 0.15). On the other hand somewhat surprisingly, we cannot reject null of the same monthly average earnings for full-time and non-full-time men on 10% significance level. This suggests that we should think of our subsequent results as being applicable to full-time men employees in the specific context but not necessarily applicable to non-full time men workers or women.

5) a) We decided to extend the model by inclusion of dummy variables for main activity of the employer. The reason for this choice is that it is likely predictive of the earnings and at the same time correlated with education (e.g. employees in manufacturing might have fewer years of education and the average wages in manufacturing can differ from the the national average). Thus omitting these variables might lead to bias. The table 2 shows the estimated coefficients for both basic and extended models. We see that while many of the newly included variables are significant, the coefficients on years of schooling and experience remain virtually unchanged which alleviates our concerns about omitted variable bias.

b) Since all the data for the survey were collected at one point in time (that is December 1996), we do not have to control for inflation in earnings. The situation would be different if we had panel data of earnings, then inflation would have to be considered otherwise the levels of earnings would not be comparable across periods.

c) To test for heteroskedasticity, we used White test. Specifically, we took squared residuals from the basic model and run the regression of them on the second polynomial of the dependent variable. The resulting LM-statistic for the joint significance of the model is 65.02 which corresponds to the $p$-value of $7.55 \cdot 10^{-15}$ (the LM-statistic is chi-squared distributed with number of restrictions as degrees of freedom). Therefore we can confidently reject the null hypothesis of homoskedasticity. For this reason, we have used heteroskedasticity robust standard errors

(in particular HC2) throughout this paper (including the basic model).

d) For any particular linear model, it is generally not possible to perform a valid statistical test for omitted variable bias without making some additional assumptions. This is because, from the definition, we generally cannot use the observed data to detect presence of some unobserved confounders.

e) Taking the derivative with respect to experience gives us return to experience as $\beta^{exp} + 2\beta^{exp\_sq} EXP$ where $\beta^{exp}$ and $\beta^{exp\_sq}$ are coefficients on experience and experience squared, respectively. This means that at 0 years of experience it is simply $\beta^{exp}$. Our estimate of $\beta^{exp}$ from the extended model is 0.02 corresponding to approximately 2% return to one year of experience for labor market entrant.

f) Using the estimates of our expended model, we get that the year of experience when maximum earnings are reached is 23.51. This can be calculated simply from taking the first order condition of the Mincer equation. Specifically, we get that

$$EXP^* = -\frac{\beta^{exp}}{2\beta^{exp\_sq}}$$

g) We estimated relationship between worker years of experience and log wages using three element connected linear splines. Figure 1 show results of estimation with fixed, evenly spaced spline nodes. Model was estimated using BFGS algorithm.
To verify robustness of our results with respect to node choice, we estimated the same model, but we allowed algorithm to optimize over the nodes (with restriction that they had to live between 0th and 90th quantile of data, to avoid problem with sparse segments). In Figure 2, can see, that broad story of our estimation doesn't change.

h) It is preferable to use variable A09 for years of education instead of A05, because A09 should more precisely estimate exact years of studying. First, we do not know if an individual attended primary school for 8 or 9 years. Second, individual may fail his study before the completion of studies, hence A05 is downward biased. Third, it may be difficult to find corresponding time of studies to attained education. Moreover, corresponding time may not be exactly determined. For example, the questionnaire provides information about apprenticeship lasting 3 to 4 years. Moreover, the information about the length of secondary vocational with GCE and grammar school with GCE is not provided (Based on our knowledge we assume additional 4 years of studying for both). Finally, it is worthy to mention disadvantages of the variable A09 too. It may be difficult for respondents to count exact years of education, hence A09 may be affected by unobserved errors.

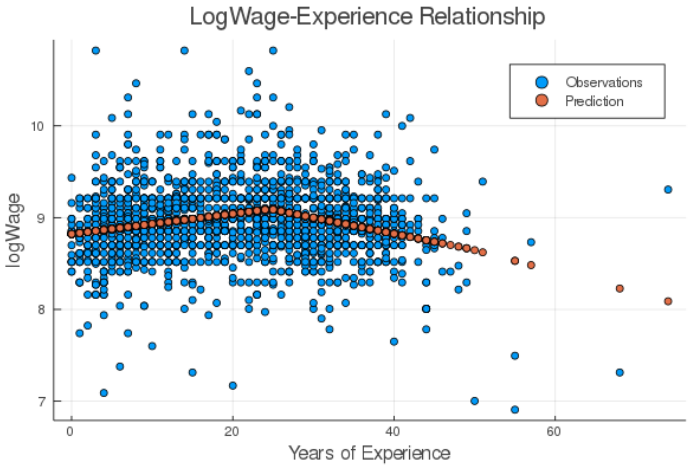


FIGURE 1. THERE IS CLEAR LIFECYCLE TREND PRESENT IN DATA. WAGES AREN'T MONOTONIC IN WORK EXPERIENCE. AFTER $\approx$ 25 YEARS OF EXPERIENCE, WAGES STARTS TO DECLINE. ECONOMICALLY, IT MAKES SENSE, TOO OLD WORKERS AREN'T MUCH PRODUCTIVE ON AVERAGE.

We can also discuss differences between both variable more quantitatively and calculate new variables Low and High education. For both variables we assume that secondary vocational and grammar schools with GCE takes 4 years. Low education account for 8 years of grammar school, 3 years of apprenticeship without DCE and 3 years of university education. High education account for 9 years of grammar school, 4 years of apprenticeship without DCE and 5 years of university education. Variable A09 has mean 12.6 and variance 2.4; variable Low education has mean 11.6 and variance 1.6; and variable High education has mean 13.3 and variance 2.1. The histograms of all three variables is provided in the graph 1.

6) The return to a year of education from the coefficients is represented in Table 3. It was calculated for each level of education in the following way (t - years of studying, b - coefficient):

$$\frac{(exp(b)-1)}{t}100$$

Note that we considered apprenticeship (3-4 years) to take 3.5 years (on average), secondary vocational with GCE and Grammar school with GCE to take 4 years, and higher education to take 5 years.

7) a) To test effect of having children on earnings, we simply created a

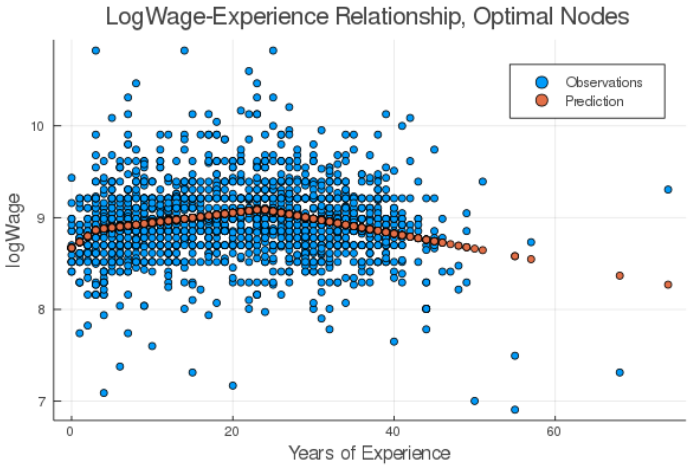


FIGURE 2. OPTIMAL SELECTION OF NODES SHOWS, THAT FIRST FEW YEARS OF WORK EXPERIENCE BRINGS HIGHEST MARGINAL INCREASES OF WAGE. MAKES SENSE WITH RESPECT TO ECONOMIC THEORY.

dummy variable *has_child* that equals 1 if the person has one or more children and zero otherwise and added it to the basic Mincer equation specification. The results are provided in table 4 (Model 2). We see that men with children have around 4% higher earnings (when keeping the education, experience and experience squared constant) although this difference is not significant at 5% level (the $p$-value for the coefficient on *has_child* is about 0.09). Note that in all cases here we will be testing hypothesis against their two-sided alternative.

b) To measure the effect of having more than one child, we include a dummy variable *more_than_one_child* into the model (while keeping *has_child*). Note that men without children are the baseline (reference) group. Model 3 of table 4 show that compared to men without children, men with exactly one child have around 5% higher earnings while men with more than one children have about 3% higher earning compared to the reference group. However standard errors for these estimates for these estimates are relatively large.

c) The null hypothesis of no differences in earnings between men with one child vs men with more then one child corresponds to equality of coefficients on *has_child* and *more_than_one_child*. This means that our null hypothesis is $H_0 : \beta^{has\_child} - \beta^{has\_child} = 0$ an we can apply Wald test with one restriction. The resulting Wald statistic is 0.33 and $p$-value of 0.57.

8) a) To allow the effect of schooling to be different in Prague, it is sufficient to add interaction of Prague dummy variable with the years of

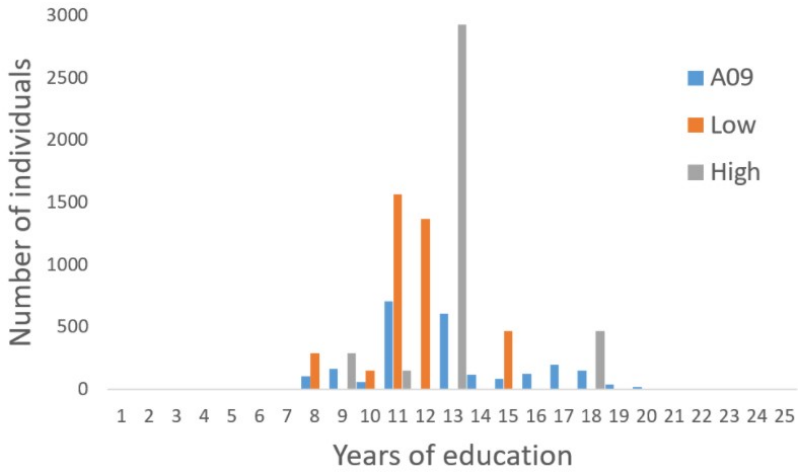


FIGURE 3. HISTOGRAMS OF EDUCATION DIRECTLY OBSERVED IN DATA (A09) AND TWO CALCULATED ALTERNATIVES (LOW AND HIGH ESTIMATES)

schooling, experience and experience squared to the model, that is

$$\begin{aligned}\log(Y_i) =&\alpha + \beta_1 EDU_i + \beta_2 EXP_i + \beta_3 EXP_i^2 + \beta_4 PRAG_i + \\ &\beta_5 PRAG_i \cdot EDU_i + \beta_6 PRAG_i \cdot EXP_i + \\ &\beta_7 PRAG_i \cdot EXP_i^2 + \epsilon_i\end{aligned} \tag{1}$$

where $PRAG_i$ is the dummy variable for Prague. The results of the specification above are provided in the table 5. Interestingly, neither the coefficient on Prague nor the interaction with schooling are significantly different from zero.

b) The null hypothesis of no effect due to Prague is equivalent to $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. The Wald statistic for this multiple restriction test is 29.63 with $p$-value of $5.18 \cdot 10^{-6}$ (since under the null Wald-statistic follows Chi-sqaured distribution with 4 degrees of freedom). This means that we can reject the null hypothesis at very high level of confidence.

c) The district unemployment has strong negative effect on earning even conditional on the controls as can be seen from results in the table 5, Model 2. The economic intuition is that the district unemployment is an (negative) indicator of the tightness of the local labor market therefore it not surprising to see it having an negative effect on earnings.

9) a) The results of the Oaxaca decomposition are provided in table 6. The Oaxaca decomposition breaks down the differences in the average logarithm of earnings between mean and women into a part that is driven

by differences in endowments (i.e. differences in average years of schooling, experience and experience squared) which is called explained variation and differences in the returns on endowments (i.e. differences in the coefficients on the independent variables for men and women). The latter part is called unexplained variation could potentially be attributed to discrimination.

b) We can test equality of the returns on education of the two genders by testing if the unexplained coefficient on schooling years (i.e. the difference in the coefficients between the genders) is significantly different from zero. We see from table 6 that we can reject the null of equal returns at 0.1% level (the coefficient is -0.316 and SE is 0.0863).

As for the hypothesis of equal levels of experience, we can look at the coefficient on *exper* in the explained part. We see that the coefficient is 0.00431 with SE of 0.0106 so it is clearly not statistically significant even at 10% level. Therefore we cannot reject the null of equality of experience levels between the genders.

c) Overall women have around 32% lower earnings than men. From this, the differences in logarithm of earnings explained by different endowments are very small and statistically insignificant. The unexplained differences (i.e. differences in return to education and experience) are much larger (in favor of men) and account for vast majority of differences in earnings between genders.

This might suggest possible discrimination of women on the labor market.

10) a) Both measures of inequality show higher concentration of earnings within men than within women. For men the Gini coefficient is 0.255 and the 90/10 decile ratio is 2.766, whereas for women the corresponding values are 0.223 and 2.571, respectively.

b) If we consider a household as s single earning unit (that is we take the average earning of every household) we get the Gini coefficient of 0.231 and the 90/10 decile ratio of 2.575.

Table 2—Basic and extended models

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 8.06*** | 7.92*** |
| | (0.06) | (0.07) |
| schooling_years | 0.06*** | 0.06*** |
| | (0.00) | (0.01) |
| exper | 0.02*** | 0.02*** |
| | (0.00) | (0.00) |
| exper_sq | −0.00*** | −0.00*** |
| | (0.00) | (0.00) |
| RawMat&EnergyDist | | 0.10* |
| | | (0.05) |
| Construction | | 0.15*** |
| | | (0.04) |
| Private Services | | 0.18*** |
| | | (0.04) |
| Public Services | | 0.07 |
| | | (0.04) |
| Finance&Ins&Real | | 0.09 |
| | | (0.08) |
| Transp&Telecom | | 0.17*** |
| | | (0.04) |
| Manuf1 | | 0.11** |
| | | (0.04) |
| Manuf2 | | 0.07 |
| | | (0.04) |
| Other | | 0.23* |
| | | (0.11) |
| $R^2$ | 0.16 | 0.18 |
| Adj. $R^2$ | 0.16 | 0.17 |
| Num. obs. | 1672 | 1672 |
| RMSE | 0.37 | 0.37 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3—Return to a year of education

| **Education level** | **Return (in %)** |
|---|---|
| Apprenticeship (2 years) | 5.11 |
| Apprenticeship (3-4 years) | 4.01 |
| Secondary vocational with GCE | 9.02 |
| Grammar school with GCE | 11.93 |
| Higher education | 14.27 |

Table 4—Effects of children

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| (Intercept) | 8.52*** | 8.50*** | 8.50*** |
| | (0.05) | (0.05) | (0.05) |
| Apprenticeship (2 years) | 0.10 | 0.10 | 0.10 |
| | (0.06) | (0.06) | (0.06) |
| Apprenticeship (3 - 4 years) | 0.13** | 0.14** | 0.14** |
| | (0.05) | (0.05) | (0.05) |
| Secondary vocational with GCE | 0.31*** | 0.31*** | 0.31*** |
| | (0.05) | (0.05) | (0.05) |
| Grammar school with GCE | 0.39*** | 0.40*** | 0.40*** |
| | (0.10) | (0.10) | (0.10) |
| Higher education | 0.54*** | 0.54*** | 0.54*** |
| | (0.06) | (0.06) | (0.06) |
| exper | 0.02*** | 0.02*** | 0.02*** |
| | (0.01) | (0.01) | (0.01) |
| exper_sq | −0.00*** | −0.00*** | −0.00*** |
| | (0.00) | (0.00) | (0.00) |
| any_children | | 0.04 | |
| | | (0.02) | |
| more_than_one_child | | | 0.03 |
| | | | (0.03) |
| one_child | | | 0.05 |
| | | | (0.03) |
| $R^2$ | 0.18 | 0.18 | 0.18 |
| Adj. $R^2$ | 0.18 | 0.18 | 0.18 |
| Num. obs. | 1672 | 1672 | 1672 |
| RMSE | 0.37 | 0.37 | 0.37 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5—Prague and regional unemployment effect

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 8.08*** | 8.14*** |
| | (0.07) | (0.06) |
| schooling_years | 0.05*** | 0.06*** |
| | (0.01) | (0.00) |
| prague | 0.05 | |
| | (0.19) | |
| exper | 0.02*** | 0.02*** |
| | (0.01) | (0.01) |
| exper_sq | −0.00** | −0.00*** |
| | (0.00) | (0.00) |
| schooling_years * prague | 0.01 | |
| | (0.01) | |
| prague * exper | 0.01 | |
| | (0.01) | |
| prague * exper_sq | −0.00 | |
| | (0.00) | |
| urate | | −2.03*** |
| | | (0.51) |
| $R^2$ | 0.18 | 0.17 |
| Adj. $R^2$ | 0.17 | 0.17 |
| Num. obs. | 1672 | 1672 |
| RMSE | 0.37 | 0.37 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6—Oaxaca decomposition

| | (1)<br>ln_y |
|---|---|
| overall | |
| Male | 8.942*** |
| | (0.0113) |
| Female | 8.626*** |
| | (0.00984) |
| difference | 0.316*** |
| | (0.0150) |
| explained | 0.000518 |
| | (0.00663) |
| unexplained | 0.315*** |
| | (0.0136) |
| explained | |
| schooling_years | 0.0194*** |
| | (0.00441) |
| exper | 0.00431 |
| | (0.0106) |
| exper_sq | -0.0232* |
| | (0.0113) |
| unexplained | |
| schooling_years | -0.333*** |
| | (0.0878) |
| exper | 0.537*** |
| | (0.128) |
| exper_sq | -0.392*** |
| | (0.0839) |
| _cons | 0.504*** |
| | (0.0943) |
| *N* | 3689 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$